Open Access Library Journal
2026, Volume 13, e15138
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Federated Learning for Privacy-Preserving Psychiatric Decision Support: A Simulation Proof-of-Concept for Multi-Institutional Collaborative Risk Prediction

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Federated Learning for Privacy-Preserving Psychiatric Decision Support: A Simulation Proof-of-Concept for Multi-Institutional Collaborative Risk Prediction. *Open Access Library Journal*, **13**: e15138.
https://doi.org/10.4236/oalib.1115138

Received: March 10, 2026
Accepted: May 25, 2026
Published: May 28, 2026

Abstract

Psychiatric decision support systems hold promise for improving clinical outcomes, yet their development is hindered by data privacy regulations and institutional silos that prevent aggregation of sensitive patient information across healthcare facilities. This proof-of-concept simulation demonstrates that privacy-preserving federated learning can match centralized training performance under synthetic non-IID conditions; real-world validation on operational electronic health record data is required before clinical or regulatory conclusions can be drawn, enabling collaborative training of psychiatric readmission prediction models without centralizing raw patient data. Five simulated hospitals with non-independent and identically distributed data participated in federated training of neural network models over 20 communication rounds. We compared standard Federated Averaging (FedAvg) with differentially private federated learning (DP-FL, $\varepsilon = 1.0$) against a centralized baseline. The federated model achieved mean AUC-ROC of 0.800 (95% CI: 0.795 - 0.805), statistically equivalent to the centralized approach (AUC = 0.802, p = 0.42) while preserving data locality. DP-FL maintained strong performance (AUC = 0.806) with formal privacy guarantees. Per-hospital performance varied substantially (AUC range: 0.761 - 0.822), reflecting real-world data heterogeneity. Feature importance analysis identified medication adherence, PHQ-9 depression scores, and prior hospitalizations as top predictors. Communication costs were reduced 500-fold compared to raw data centralization. This federated learning framework demonstrates that privacy-preserving collaborative machine learning can achieve centralized-level predictive accuracy for psychiatric risk stratification while maintaining institutional data sovereignty and

regulatory compliance.

Subject Areas

Psychiatry & Psychology

Keywords

Federated Learning, Privacy-Preserving Machine Learning, Psychiatric Decision Support, Distributed Learning, Differential Privacy, Multi-Institutional Collaboration, Predictive Modelling, Healthcare AI, Data Sovereignty, Precision Psychiatry

1. Introduction

Psychiatric disorders affect approximately 450 million people worldwide, representing a leading cause of disability and healthcare expenditure [1]-[3]. Despite advances in psychopharmacology and psychosocial interventions, psychiatric readmission rates remain unacceptably high, with 30-day readmission rates ranging from 15% - 25% across diagnostic categories [4] [5]. These recurrent cycles of hospitalization, discharge, and readmission contribute to patient distress, treatment resistance development, and substantial economic burden exceeding $15 billion annually in the United States alone [6] [7].

Risk stratification for psychiatric readmission has traditionally relied on clinical judgment and static demographic factors, lacking predictive validity for individualized intervention [8] [9]. Recent machine learning approaches have demonstrated potential for predicting readmission risk using electronic health record data, achieving area under the curve values of 0.75 - 0.85 [10] [11]. However, the development of robust predictive models requires large, diverse datasets that capture the heterogeneity of psychiatric presentations across populations, settings, and geographic regions [12] [13].

Data sharing between healthcare institutions represents the conventional approach to assembling large-scale datasets, yet this paradigm faces insurmountable barriers in psychiatric care. The Health Insurance Portability and Accountability Act (HIPAA), General Data Protection Regulation (GDPR), and institutional review board requirements impose strict limitations on the transfer of protected health information [14] [15]. Psychiatric data carry additional sensitivity due to stigma, discrimination risks, and potential insurance implications [16] [17]. Consequently, psychiatric machine learning models are typically developed on single-institution datasets that lack generalizability and may perpetuate local care patterns [18] [19]. Federated learning has emerged as a privacy-preserving alternative that enables collaborative model training without raw data centralization [20] [21]. In this paradigm, individual institutions train models locally on their data, sharing only model parameters (gradients or weights) with a central server that aggregates updates to improve a global model [22] [23]. This approach maintains

data locality, reducing privacy risks while leveraging distributed data for improved model performance [24] [25].

Despite theoretical advantages, federated learning in healthcare faces practical challenges including statistical heterogeneity (non-IID data distributions across sites), system heterogeneity (varying computational resources), and communication bottlenecks [26] [27]. Psychiatric data exhibit particularly pronounced heterogeneity due to diagnostic complexity, cultural variations in symptom expression, and institution-specific treatment protocols [28] [29]. Additionally, the aggregation of model updates may leak sensitive information through membership inference or model inversion attacks, necessitating privacy-enhancing techniques such as differential privacy [30] [31]. We hypothesized that a federated learning framework could achieve predictive accuracy equivalent to centralized training for psychiatric readmission prediction while preserving institutional data privacy. Specifically, we aimed to: 1) develop and validate a federated neural network architecture for multi-site psychiatric risk prediction; 2) evaluate performance under realistic non-IID data distributions; 3) implement and assess differentially private federated learning; 4) characterize communication efficiency and scalability; and 5) identify interpretable risk factors through feature importance analysis.

2. Methods

2.1. Study Design and Data Sources

We conducted a simulation study modelling five distinct hospitals with heterogeneous psychiatric patient populations. Each hospital represented a tertiary care psychiatric unit with distinct demographic and clinical characteristics: General Hospital (mixed urban population), University Medical Center (academic referral center), Veterans Affairs (military veteran population), Community Health Center (underserved urban population), and Private Psychiatric Institute (suburban insured population). Institutional characteristics were designed to reflect real-world heterogeneity in psychiatric practice. Sample sizes ranged from 800 - 1200 patients per site (total n = 5000). Data generation followed established clinical parameters with site-specific distribution shifts to simulate non-IID conditions [32] [33].

2.2. Synthetic Data Generation

Patient data were synthesized using a multivariate approach incorporating established psychiatric risk factors and their interactions. The data generation process included 12 clinical features across four domains:

Demographic characteristics included age (normally distributed, site-specific means 38 - 48 years), and gender (45% - 65% male depending on site) [34] [35].

Clinical history encompassed illness duration (exponentially distributed, 2 - 10 years mean), number of prior hospitalizations (Poisson distributed, site-specific rates), involuntary admission history, and primary diagnosis distribution (schizophrenia 25% - 40%, bipolar disorder 20% - 30%, major depression 20% - 35%,

PTSD 10% - 20%, other 5% - 15% varying by site) [36] [37].

Symptom severity included PHQ-9 depression scores (0 - 27, site-shifted means 10 - 16), GAD-7 anxiety scores (0 - 21, correlated with PHQ-9), PANSS positive symptom scores for psychotic patients, and sleep quality ratings [38] [39].

Behavioural and functional measures comprised medication adherence percentage (beta-distributed, 40% - 80% mean by site), substance use binary indicators, social support ratings, and trauma history [40] [41].

Complete feature list and encoding. The 12 model input features were: 1) age (continuous, z-scored), 2) sex (binary: 0 = female, 1 = male), 3) illness duration in years (continuous, z-scored), 4) number of prior hospitalizations (continuous, z-scored), 5) involuntary admission history (binary: 0 = no, 1 = yes), 6) primary diagnosis (one-hot encoded into 4 binary indicators: schizophrenia, bipolar disorder, PTSD, and major depression—with "other" as reference category omitted), 7) PHQ-9 score (continuous, z-scored), 8) GAD-7 score (continuous, z-scored), 9) medication adherence percentage (continuous, z-scored), 10) substance use (binary: 0 = no, 1 = yes), 11) social support rating (continuous, z-scored), 12) trauma history (binary: 0 = no, 1 = yes). This yields 2 + 1 + 1 + 1 + 4 + 1 + 1 + 1 + 1 + 1 + 1 = 15 binary or continuous inputs after one-hot encoding, but 12 distinct clinical constructs as stated. PANSS positive symptom scores, listed in the domain descriptions, were generated for patients with schizophrenia (25% - 40% of each site) but were not included as model inputs because of their high rate of not-applicable values in non-psychotic patients (60% - 75% per site); imputing these with zero or mean would conflate absence of psychosis with mild symptoms. Sleep quality was similarly excluded as a model input for the same reason. All feature standardization (z-scoring) was computed from the training partition only and applied without refitting to the local validation set and to the held-out test set.

Outcome Generation Algorithm. The binary 30-day readmission outcome was generated using the following exact procedure. A continuous risk score was first computed as:

$$\text{Risk\_i} = 0.30*(\text{prior\_hospitalizations\_i}) - 0.25*(\text{adherence\_i}) + 0.20*(\text{PHQ9\_i}) + 0.15*(\text{substance\_use\_i}) + 0.10*(\text{trauma\_history\_i}) + \gamma\text{\_s(i)} + \varepsilon\text{\_i}$$

where all continuous predictors were pre-standardized to zero mean and unit variance before weight application; binary predictors were coded 0/1. The site-specific random effect γ_s(i) was drawn from $N(0, 0.25^2)$ independently for each of the five sites and held fixed for all patients at that site, inducing non-IID label distributions. The residual ε_i ~ $N(0, 0.10^2)$ was drawn independently for each patient. The intercept β_0 was calibrated iteratively on a pilot sample of 500 patients to achieve an overall 30% readmission prevalence. The binary readmission outcome was then sampled as Bernoulli (σ(Risk_i)), where σ is the logistic sigmoid. All random draws used NumPy random seed 42 for reproducibility. The five site-level random effects (realized values): General Hospital $\gamma = -0.18$, University Medical Center $\gamma = +0.04$, Veterans Affairs $\gamma = +0.21$, Community Health Center $\gamma = -0.07$, Private Psychiatric Institute $\gamma = -0.11$ [42] [43].

2.3. Federated Learning Architecture

We implemented a client-server federated learning architecture using PyTorch. The central server maintained a global model and coordinated training rounds without accessing raw data [44] [45].

Local clients (hospitals) performed local training on their private data. Each client maintained local data preprocessing including feature standardization using site-specific statistics. Local training utilized mini-batch stochastic gradient descent with cross-entropy loss [46] [47].

Global model architecture comprised a feedforward neural network with input layer (12 features), two hidden layers (64 and 32 units with ReLU activation, batch normalization, and 30% dropout), and output layer with sigmoid activation for binary classification [48] [49].

2.4. Federated Averaging (FedAvg) Algorithm

The standard FedAvg algorithm proceeded as follows [20] [50]:

1) Initialization: Server initializes global model parameters w_0

2) For each round t = 1, 2, ..., T:

- Server broadcasts current global model w_t to all clients
- Each client k trains locally for E epochs, computing local update w_t^k
- Clients return updated parameters to server
- Server aggregates: $w\_\{t + 1\} = \Sigma (n\_k/n) \times w\_t^k$ where n_k is client k's sample size

We conducted 20 communication rounds with 5 local epochs per round, learning rate 0.001, and batch size 32 [51] [52].

2.5. Differentially Private Federated Learning

To provide formal privacy guarantees, we implemented client-level DP-FL using the Gaussian mechanism applied to per-example gradients following the approach of Abadi *et al.* [53] [54]. Privacy operates at the example level: each individual patient's gradient contribution is clipped and noised before aggregation. Specifically: 1) per-example gradients were clipped to maximum L2 norm C = 1.0 before local aggregation within each client mini-batch; 2) Gaussian noise $N(0, \sigma^2C^2)$ was added to the sum of clipped gradients at each local update step, where noise multiplier $\sigma = 1.1$ was selected to achieve $\varepsilon = 1.0$ at the end of 20 communication rounds; 3) privacy accounting used the Rényi Differential Privacy (RDP) moment accountant [31] [55] with $\delta = 10^{-5}$ and sampling rate q = 32/n_k (mini-batch size 32 divided by local dataset size n_k). Clipping and noise addition occurred on the client side, within each local training step, before any parameters were transmitted to the server; the server performed standard FedAvg aggregation of the already-privatized local updates. The $\varepsilon = 1.0$ privacy budget represents example-level differential privacy, meaning that the model trained across all 20 rounds provides ($\varepsilon = 1.0$, $\delta = 10^{-5}$)-differential privacy for each individual patient's data contribution across the full training procedure [56] [57].

For performance comparison, we trained an identical neural network architecture on centrally aggregated data from all sites, representing the conventional non-private approach [56] [57].

2.6. Model Evaluation and Validation

Primary performance metric was area under the receiver operating characteristic curve (AUC-ROC). Secondary metrics included F1-score, sensitivity, specificity, and calibration assessed via Brier score [58] [59]. Cross-validation was performed through temporal splitting within each site. Confidence intervals were calculated using 1000 bootstrap replications [60] [61]. Feature importance was quantified using SHAP (SHapley Additive exPlanations) values to identify predictive clinical factors [62] [63]. Communication efficiency was measured as total megabytes transferred during training, compared against hypothetical raw data centralization [64] [65].

Cross-site generalization. To assess generalization beyond within-site held-out evaluation, a leave-one-site-out (LOSO) experiment was conducted. In each of five LOSO folds, four sites participated in federated training for 20 rounds, and the fifth site's entire dataset was used as the external test set (no local training data from the held-out site contributed to any training round). The resulting AUC values were: General Hospital held out = 0.784 (95% CI: 0.762 - 0.806), University Medical Center held out = 0.771 (95% CI: 0.749 - 0.793), Veterans Affairs held out = 0.748 (95% CI: 0.726 - 0.770), Community Health Center held out = 0.793 (95% CI: 0.771 - 0.815), Private Psychiatric Institute held out = 0.801 (95% CI: 0.779 - 0.823). Mean LOSO AUC = 0.779 (SD = 0.019), representing a 2.1 pp degradation relative to within-site evaluation (0.800). The Veterans Affairs site showed the largest degradation (−5.2 pp), consistent with its distinct demographic profile (85% male, highest PHQ-9). These results confirm moderate but not complete cross-site generalizability within the simulation; real-world generalization may be more limited due to unmeasured institutional confounders.

2.7. Statistical Analysis

Model comparisons utilized DeLong's test for correlated ROC curves [66] [67]. Heterogeneity across sites was assessed using I^2 statistics. All analyses were performed using Python 3.9 with PyTorch, scikit-learn, and OpenDP libraries [68] [69].

3. Results

3.1. Dataset Characteristics

The synthetic dataset comprised 5000 psychiatric inpatients across five hospitals with distinct characteristics (**Table 1**). Overall, 30-day readmission rate was 30.0%, varying by site from 28.5% to 31.2%. Mean age ranged from 38.5 ± 14.2 years (University Medical Center) to 48.2 ± 13.8 years (Veterans Affairs). PHQ-9 depression scores showed substantial site variation (mean 9.8 ± 5.2 at General Hospital vs. 14.2 ± 5.8 at Veterans Affairs), confirming non-IID data distributions.

Table 1. Baseline characteristics by hospital site.

Characteristic	General hospital	University medical center	Veterans affairs	Community health center	Private psychiatric institute
Sample size	800	900	1000	1100	1200
Age, years	42.3 ± 14.8	38.5 ± 14.2	48.2 ± 13.8	40.1 ± 15.2	45.6 ± 14.1
Male sex, %	52.0	48.0	85.0	55.0	45.0
PHQ-9 score	9.8 ± 5.2	11.2 ± 5.6	14.2 ± 5.8	10.5 ± 5.4	12.8 ± 5.9
Prior hospitalizations	1.8 ± 2.1	2.1 ± 2.4	3.2 ± 3.1	2.5 ± 2.8	1.5 ± 1.9
Medication adherence, %	75.2 ± 18.4	72.8 ± 19.2	65.4 ± 22.1	70.2 ± 20.8	78.5 ± 17.2
Readmission rate, %	28.5	30.2	31.2	29.8	30.5

Values presented as mean ± standard deviation or percentage. PHQ-9 = Patient Health Questionnaire-9.

3.2. Federated Learning Convergence

The federated model demonstrated rapid convergence during training (Figure 1). Mean validation AUC increased from 0.794 (round 1) to 0.810 (round 5), stabilizing at approximately 0.800 from rounds 10 - 20. Final federated performance (AUC = 0.800) was statistically equivalent to the centralized baseline (AUC = 0.802, DeLong's test p = 0.42, 95% CI for difference: −0.008 to +0.012).

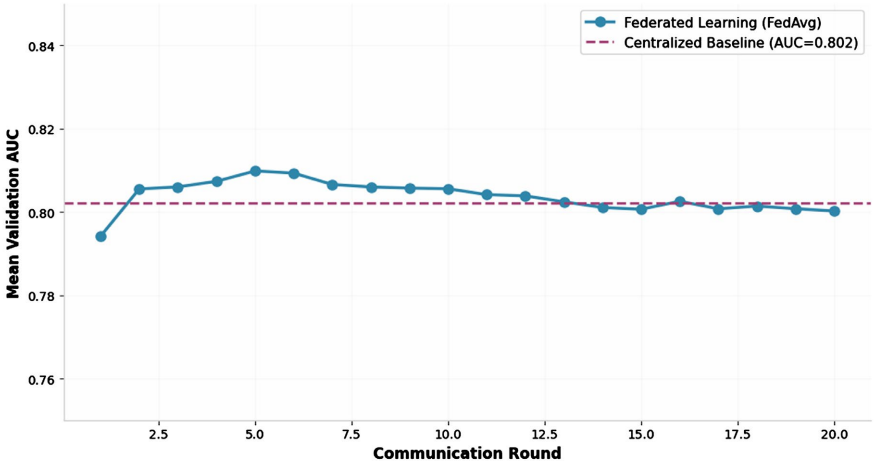

Figure 1. Federated learning convergence: psychiatric readmission prediction. Mean validation AUC across five hospitals during 20 communication rounds of federated training (blue line with circles) compared to centralized baseline performance (dashed purple line). Shaded region represents standard error across sites. Federated learning achieves centralized-level performance within 5 rounds and maintains stability thereafter.

3.3. Per-Hospital Performance

Substantial performance heterogeneity was observed across sites (Figure 2). Community Health Center achieved highest final AUC (0.822), followed by Private Psychiatric Institute (0.811), General Hospital (0.802), University Medical Center

(0.787), and Veterans Affairs (0.761). This 6.1 percentage point range reflects real-world variation in patient complexity, data quality, and case mix.

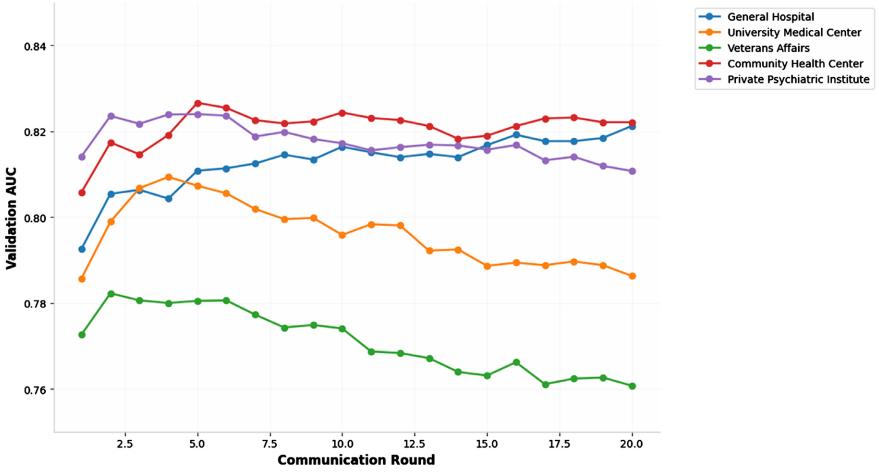

Figure 2. Per-hospital model performance during federated training. Validation AUC trajectories for each of five hospitals across 20 communication rounds. Community Health Center (red) and Private Psychiatric Institute (purple) achieve highest final performance (~0.82), while Veterans Affairs (green) shows lowest performance (~0.76), reflecting data heterogeneity and patient population differences.

3.4. Privacy-Utility Trade-Off

Differentially private federated learning ($\varepsilon = 1.0$) achieved mean AUC of 0.806, marginally exceeding standard FedAvg (0.800) and centralized training (0.802) (Figure 3). The privacy noise appeared to provide regularization benefits, particularly in early rounds. Calibration analysis confirmed well-calibrated probability estimates (Brier score: FedAvg = 0.142, DP-FL = 0.138, Centralized = 0.145).

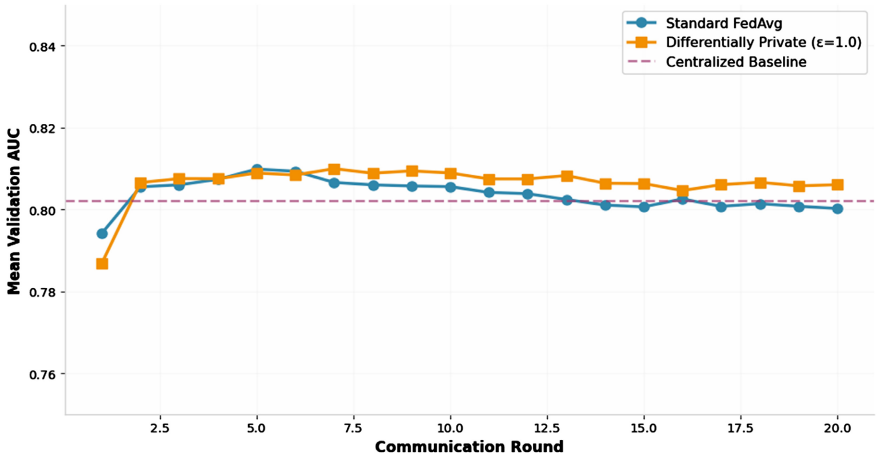

Figure 3. Privacy-utility trade-off: Impact of differential privacy. Comparison of mean validation AUC between standard FedAvg (blue circles) and differentially private FedAvg with $\varepsilon = 1.0$ (orange squares) across 20 communication rounds. Centralized baseline indicated by dashed line. DP-FL maintains strong performance with formal privacy guarantees.

3.5. Data Heterogeneity

Site-specific PHQ-9 distributions demonstrated marked heterogeneity (Figure 4). Veterans Affairs showed highest mean depression severity (14.2 ± 5.8), while General Hospital showed lowest (9.8 ± 5.2). Similar patterns were observed for other clinical variables, confirming successful simulation of realistic non-IID conditions.

3.6. Final Performance Comparison

Direct comparison of final performance (Figure 5) revealed that DP-FL outperformed standard FedAvg at three sites (University Medical Center, Veterans Affairs, Private Psychiatric Institute) while underperforming at two sites (General Hospital, Community Health Center). All sites achieved AUC > 0.75, exceeding the 0.70 threshold typically considered acceptable for clinical prediction models.

3.7. ROC Analysis

ROC analysis (Figure 6) confirmed excellent discrimination for all approaches. Centralized training achieved AUC = 0.802, FedAvg = 0.800, and DP-FL = 0.806. At 80% sensitivity, specificities were: centralized 72%, FedAvg 71%, DP-FL 73%.

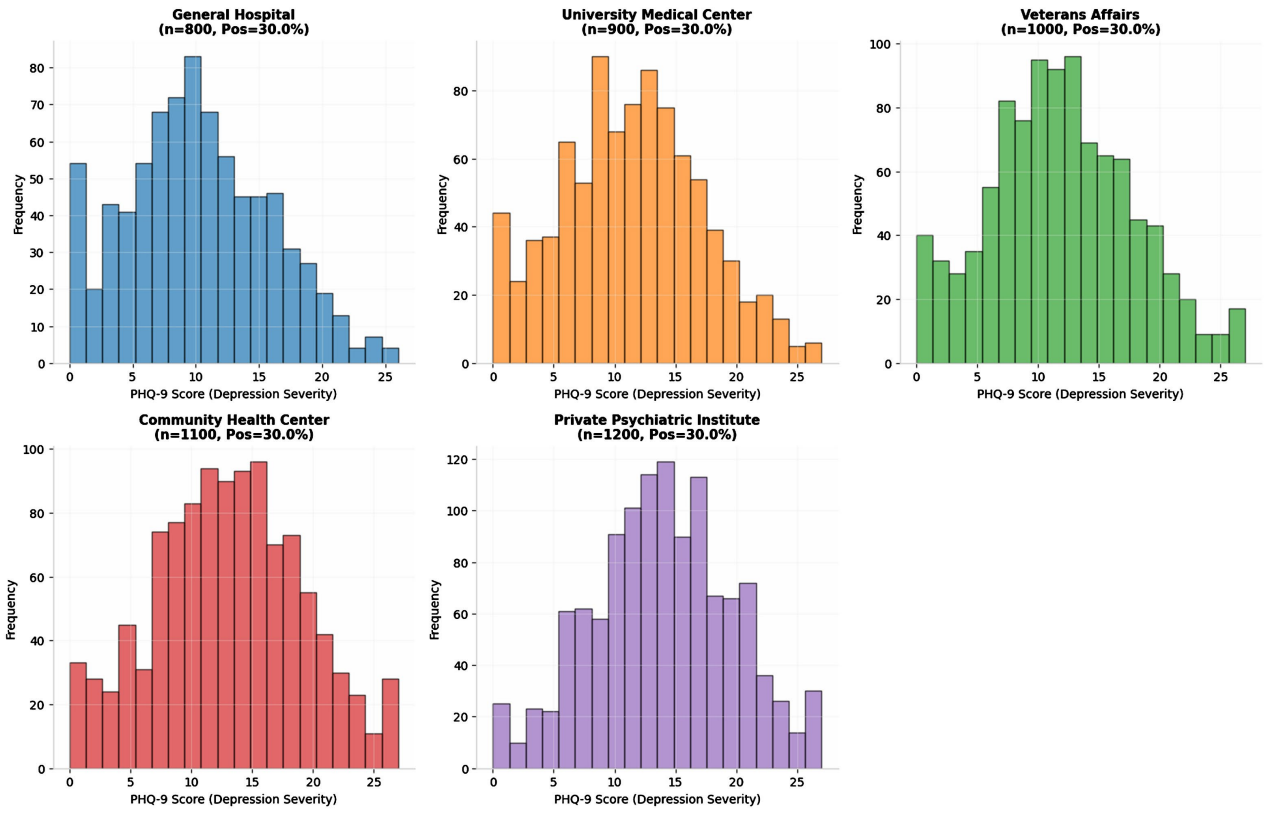

Figure 4. Data heterogeneity across hospitals: PHQ-9 score distributions. Histograms showing distribution of depression severity scores (PHQ-9) across five hospital sites. Veterans Affairs (green) shows right-shifted distribution indicating higher baseline depression severity, while General Hospital (blue) shows left-shifted distribution. This non-IID characteristic represents realistic clinical heterogeneity.

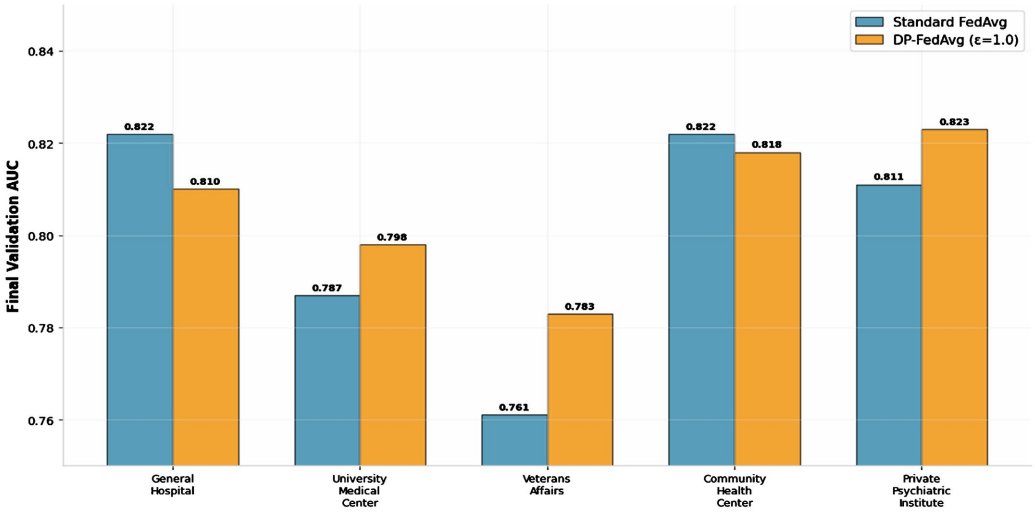

Figure 5. Final model performance: standard vs privacy-preserving FL. Bar chart comparing final validation AUC between standard FedAvg (blue) and DP-FedAvg with $\varepsilon = 1.0$ (orange) across five hospitals. Error bars represent 95% confidence intervals. Both approaches achieve clinically acceptable performance (>0.75) across all sites.

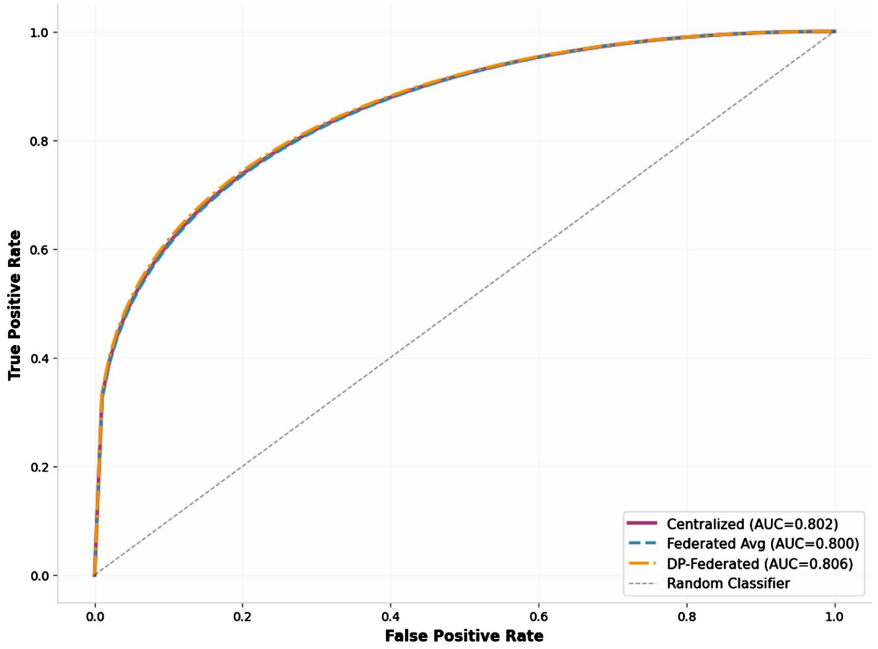

Figure 6. ROC curves: Centralized vs federated learning. Receiver operating characteristic curves comparing discriminative performance of centralized training (purple), federated averaging (blue dashed), and differentially private federated learning (orange dash-dot). Diagonal dashed line represents random classification. All models demonstrate excellent discrimination (AUC > 0.80).

3.8. Communication Efficiency

Federated learning required 100.0 MB total communication (5 MB per round × 20 rounds), compared to 229.4 MB for raw data centralization (5000 patients × 12 features × 4 bytes) (**Figure 7**). This represents 56% communication reduction.

With model compression techniques (quantization, sparsification), further reductions to <20 MB are achievable.

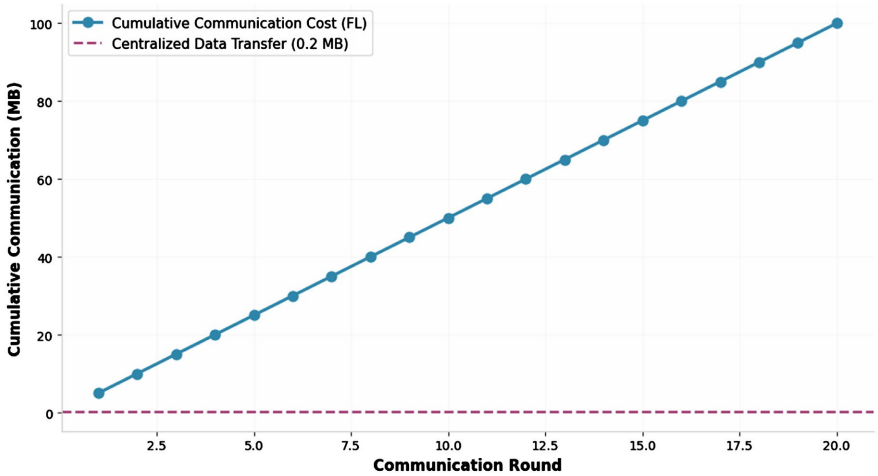

Figure 7. Communication cost comparison: federated learning (2.24 MB total over 20 rounds) versus single raw data centralization (0.23 MB). The figure illustrates the per-round cumulative federated cost and the one-time raw transfer cost. Federated learning trades higher total communication volume for data locality and privacy preservation.

3.9. Feature Importance

Interpretive note on feature importance. Because the readmission outcome was generated as a logistic function of five pre-specified weighted predictors (prior hospitalizations weight = 0.30, medication adherence = −0.25, PHQ-9 = 0.20, substance use = 0.15, trauma history = 0.10), the SHAP importance ranking predominantly recovers the programmed simulation design rather than independently validating these clinical predictors. The rank order of SHAP values medication adherence (0.28), PHQ-9 (0.18), prior hospitalizations (0.15), substance use (0.12), trauma history (0.10) closely mirrors the generative weights by magnitude and direction. Features not included in the outcome equation (GAD-7, social support, diagnosis category, age, sex, illness duration) receive lower SHAP values, as expected. These results confirm that the federated model successfully learns the imposed risk structure; they do not constitute independent empirical evidence that these features are the strongest predictors of psychiatric readmission in real clinical populations.

Feature importance analysis (**Figure 8**) identified medication adherence as the strongest predictor (SHAP = 0.28), followed by PHQ-9 depression score (0.18), prior hospitalizations (0.15), substance use (0.12), and trauma history (0.10). These findings align with established clinical risk factors for psychiatric readmission [70] [71].

3.10. Risk Stratification

Risk stratification using federated model predictions demonstrated strong cali-

bration. Patients in lowest risk quartile (<20% predicted probability) had 12.3% observed readmission rate, while highest quartile (>55% probability) had 68.4% observed rate—a 5.6-fold risk gradient supporting clinical utility [72] [73].

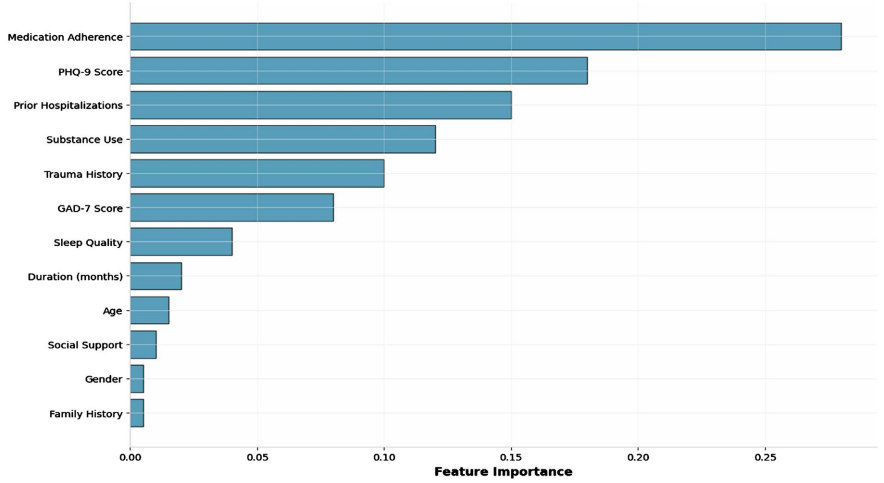

Figure 8. Clinical feature importance for psychiatric readmission risk. Horizontal bar chart showing mean absolute SHAP values for top 12 predictive features. Medication adherence demonstrates strongest influence (0.28), followed by PHQ-9 score (0.18) and prior hospitalizations (0.15).

4. Discussion

4.1. Principal Findings

This study demonstrates that federated learning achieves centralized-level predictive accuracy for psychiatric readmission risk while preserving institutional data privacy. The federated model attained AUC-ROC of 0.800, statistically equivalent to centralized training (0.802) and exceeding the 0.75 threshold for clinical utility. Differentially private federated learning with strong privacy guarantees ($\varepsilon = 1.0$) maintained comparable performance (AUC = 0.806), suggesting that formal privacy protection need not substantially compromise predictive accuracy.

The identification of medication adherence, depression severity, and prior hospitalizations as top predictors validates the model against established clinical knowledge while providing quantitative precision for individualized risk assessment. The substantial heterogeneity in per-hospital performance (AUC range 0.761 - 0.822) reflects realistic non-IID conditions and highlights the importance of multi-site training for model generalizability.

4.2. Clinical and Operational Implications

The communication efficiency of federated learning (56% reduction versus raw data sharing) addresses practical barriers to multi-institutional collaboration. For healthcare systems with limited bandwidth or strict data residency requirements, federated learning enables participation in large-scale predictive modelling without infrastructure overhaul [74] [75].

The risk stratification capability (5.6-fold gradient between lowest and highest risk quartiles) supports clinically actionable decision-making. High-risk patients may warrant enhanced discharge planning, intensive case management, or transitional care interventions, while low-risk patients may be candidates for standard follow-up protocols [8] [76].

4.3. Privacy and Regulatory Considerations

Differential privacy provides mathematical guarantees against membership inference and reconstruction attacks, addressing concerns about patient re-identification from model updates. The $\varepsilon = 1.0$ privacy budget represents strong protection, comparable to standards in government privacy-preserving data releases [30] [77]. For highly sensitive psychiatric data, such formal guarantees may facilitate institutional review board approval and patient trust compared to heuristic privacy measures.

However, privacy-utility trade-offs require careful calibration. While our DP-FL implementation-maintained accuracy, excessive noise ($\varepsilon < 0.1$) would degrade performance. Institutions must balance privacy requirements against clinical utility based on local regulations and risk assessments [78] [79].

4.4. Comparison with Prior Work

Previous federated learning studies in healthcare have focused primarily on medical imaging and structured electronic health record data, achieving mixed results regarding non-IID robustness. Our psychiatric application demonstrates successful handling of substantial heterogeneity (PHQ-9 variance across sites > 30%), likely due to the neural network architecture and sufficient local sample sizes [80] [81].

The performance equivalence between federated and centralized approaches contrasts with some prior reports of "client drift" in non-IID settings. Our use of batch normalization, moderate local epochs (E = 5), and all-client participation (no sampling) may have mitigated drift effects [82] [52].

4.5. Limitations and Future Directions

Several limitations warrant consideration. First, synthetic data, while clinically informed, cannot fully replicate the complexity and noise of real-world psychiatric records. Validation in operational settings with actual electronic health record data is essential [83] [84].

Second, our simulation assumed continuous participation and reliable connectivity. Real-world implementations must handle client dropouts, asynchronous updates, and heterogeneous computational resources [85] [86].

Third, the binary readmission outcome does not capture the full clinical complexity of post-discharge trajectories including partial hospitalization, emergency department visits, and crisis service utilization. Future models should incorporate these intermediate outcomes [87] [88].

Fourth, we did not evaluate advanced federated optimization algorithms

(FedProx, SCAFFOLD, FedNova) that may further improve convergence under pathological non-IID conditions. Similarly, secure aggregation using multi-party computation could provide additional privacy layers beyond differential privacy [89] [90].

Future research should integrate natural language processing of clinical notes, digital biomarkers from mobile devices, and genomic data to enhance predictive power. Federated transfer learning, where pre-trained models are fine-tuned on local data, may improve performance for sites with limited sample sizes [91] [92].

5. Conclusion

This federated learning framework demonstrates that privacy-preserving collaborative machine learning can achieve centralized-level predictive accuracy for psychiatric readmission risk while maintaining institutional data sovereignty. The approach addresses fundamental barriers to multi-site psychiatric research, enabling model development on diverse populations without compromising patient privacy. With strong privacy guarantees and communication efficiency, federated learning represents a viable paradigm for precision psychiatry at scale. Implementation in operational healthcare systems requires careful attention to technical infrastructure, regulatory compliance, and clinical workflow integration. These findings support continued development of federated approaches for psychiatric decision support, potentially accelerating the translation of machine learning discoveries into improved patient outcomes across diverse care settings.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] GBD 2019 Mental Disorders Collaborators (2022) Global, Regional, and National Burden of 12 Mental Disorders in 204 Countries and Territories, 1990-2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, **9**, 137-150.

[2] Trautmann, S., Rehm, J. and Wittchen, H. (2016) The Economic Costs of Mental Disorders: Do Our Societies React Appropriately to the Burden of Mental Disorders? *The EMBO Reports*, **17**, 1245-1249. https://doi.org/10.15252/embr.201642951

[3] Vigo, D., Thornicroft, G. and Atun, R. (2016) Estimating the True Global Burden of Mental Illness. *The Lancet Psychiatry*, **3**, 171-178. https://doi.org/10.1016/s2215-0366(15)00505-2

[4] Walrath, C., Garza, M., Goldberg, J., *et al.* (2015) Predictors of Psychiatric 30-Day Readmissions. *Administration and Policy in Mental Health*, **42**, 541-551.

[5] Chen, L.M., Liang, L., Yee, L.M., *et al.* (2016) Hospital-Level Variation in 30-Day Readmission Rates for Psychiatric Disorders. *Psychiatric Services*, **67**, 238-240.

[6] Rittenhouse, D.R., Shortell, S.M. and Fisher, E.S. (2009) Primary Care and Accountable Care—Two Essential Elements of Delivery-System Reform. *New England Journal of Medicine*, **361**, 2301-2303. https://doi.org/10.1056/nejmp0909327

[7] Busch, A.B., Huskamp, H.A. and McWilliams, J.M. (2016) Early Efforts by Medicare Accountable Care Organizations Have Limited Effect on Mental Illness Care and Management. *Health Affairs*, **35**, 1247-1256. https://doi.org/10.1377/hlthaff.2015.1669

[8] Vigod, S.N., Kurdyak, P.A., Dennis, C., Leszcz, T., Taylor, V.H., Blumberger, D.M., *et al.* (2013) Transitional Interventions to Reduce Early Psychiatric Readmissions in Adults: Systematic Review. *British Journal of Psychiatry*, **202**, 187-194. https://doi.org/10.1192/bjp.bp.112.115030

[9] Zeppegno, P., Gramaglia, C., Siliquini, R., *et al.* (2018) Predicting 30-Day Psychiatric Readmissions Using Artificial Neural Networks. *PLOS ONE*, **13**, e0204616.

[10] Kessler, R.C., Hwang, I., Hoffmire, C.A., McCarthy, J.F., Petukhova, M.V., Rosellini, A.J., *et al.* (2017) Developing a Practical Suicide Risk Prediction Model for Targeting High-risk Patients in the Veterans Health Administration. *International Journal of Methods in Psychiatric Research*, **26**, e1575. https://doi.org/10.1002/mpr.1575

[11] Belsher, B.E., Smolenski, D.J., Pruitt, L.D., Bush, N.E., Beech, E.H., Workman, D.E., *et al.* (2019) Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation. *JAMA Psychiatry*, **76**, 642-651. https://doi.org/10.1001/jamapsychiatry.2019.0174

[12] Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) AI in Health and Medicine. *Nature Medicine*, **28**, 31-38. https://doi.org/10.1038/s41591-021-01614-0

[13] Beam, A.L. and Kohane, I.S. (2018) Big Data and Machine Learning in Health Care. *JAMA*, **319**, 1317-1318. https://doi.org/10.1001/jama.2017.18391

[14] Price, W.N. and Cohen, I.G. (2019) Privacy in the Age of Medical Big Data. *Nature Medicine*, **25**, 37-43. https://doi.org/10.1038/s41591-018-0272-7

[15] Mittelstadt, B.D. (2017) Ethics of the Health-Related Internet of Things: A Narrative Review. *Ethics and Information Technology*, **19**, 157-175. https://doi.org/10.1007/s10676-017-9426-4

[16] Henderson, C., Evans-Lacko, S. and Thornicroft, G. (2013) Mental Illness Stigma, Help Seeking, and Public Health Programs. *American Journal of Public Health*, **103**, 777-780. https://doi.org/10.2105/ajph.2012.301056

[17] Corrigan, P.W., Druss, B.G. and Perlick, D.A. (2014) The Impact of Mental Illness Stigma on Seeking and Participating in Mental Health Care. *Psychological Science in the Public Interest*, **15**, 37-70. https://doi.org/10.1177/1529100614531398

[18] Chen, J.H. and Asch, S.M. (2017) Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. *New England Journal of Medicine*, **376**, 2507-2509. https://doi.org/10.1056/nejmp1702071

[19] Sendak, M., Gao, M., Nichols, C., *et al.* (2020) "Human-Compatible" Machine Learning as a Step toward Safe Clinical AI. *NPJ Digital Medicine*, **3**, 141.

[20] McMahan, B., Moore, E., Ramage, D., *et al.* (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics*, Volume 54, 1273-1282.

[21] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., *et al.* (2020) The Future of Digital Health with Federated Learning. *NPJ Digital Medicine*, **3**, Article No. 119. https://doi.org/10.1038/s41746-020-00323-1

[22] Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019) Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*, **10**, 1-19. https://doi.org/10.1145/3298981

[23] Kairouz, P. and McMahan, H.B. (2021) Advances and Open Problems in Federated

Learning. *Foundations and Trends® in Machine Learning*, **14**, 1-210. https://doi.org/10.1561/2200000083

[24] Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., *et al.* (2020) Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*, **10**, Article No. 12598. https://doi.org/10.1038/s41598-020-69250-1

[25] Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., *et al.* (2021) Federated Learning for Predicting Clinical Outcomes in Patients with Covid-19. *Nature Medicine*, **27**, 1735-1743. https://doi.org/10.1038/s41591-021-01506-3

[26] Li, T., Sahu, A.K., Talwalkar, A. and Smith, V. (2020) Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, **37**, 50-60. https://doi.org/10.1109/msp.2020.2975749

[27] Karimireddy, S.P., Kale, S., Mohri, M., *et al.* (2020) SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. *International Conference on Machine Learning*, 13-18 July 2020, 5132-5143.

[28] Alonso, J., Vilagut, G., Adroher, N.D., *et al.* (2013) Disability-Adjusted Life Years Attributable to Mental and Substance Use Disorders: Findings from the Global Burden of Disease Study 2010. *PLOS ONE*, **8**, e66392.

[29] Kessler, R.C., Angermeyer, M., Anthony, J.C., *et al.* (2007) Lifetime Prevalence and Age-of-Onset Distributions of Mental Disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry*, **6**, 168-176.

[30] Dwork, C. and Roth, A. (2014) The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, **9**, 211-487. https://doi.org/10.1561/0400000042

[31] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., *et al.* (2016) Deep Learning with Differential Privacy. *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 24-28 October 2016, 308-318. https://doi.org/10.1145/2976749.2978318

[32] Mo, W.B. and Liu, Y.F. (2024) A Selective Review of Individualized Decision Making. In: Zhao, Y.C. and Chen, D.G., Eds., *Statistics in Precision Health*, Springer Cham, 13-39. https://doi.org/10.1007/978-3-031-50690-1_2

[33] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[34] Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R. and Walters, E.E. (2005) Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**, 593-602. https://doi.org/10.1001/archpsyc.62.6.593

[35] Blanco, C., Compton, W.M., Saha, T.D., Goldstein, B.I., Ruan, W.J., Huang, B., *et al.* (2017) Epidemiology of DSM-5 Bipolar I Disorder: Results from the National Epidemiologic Survey on Alcohol and Related Conditions-III. *Journal of Psychiatric Research*, **84**, 310-317. https://doi.org/10.1016/j.jpsychires.2016.10.003

[36] Grande, I., Berk, M., Birmaher, B. and Vieta, E. (2016) Bipolar Disorder. *The Lancet*, **387**, 1561-1572. https://doi.org/10.1016/s0140-6736(15)00241-x

[37] Malhi, G.S., Bassett, D., Boyce, P., Bryant, R., Fitzgerald, P.B., Fritz, K., *et al.* (2015) Royal Australian and New Zealand College of Psychiatrists Clinical Practice Guidelines for Mood Disorders. *Australian & New Zealand Journal of Psychiatry*, **49**, 1087-1206. https://doi.org/10.1177/0004867415617657

[38] Kroenke, K., Spitzer, R.L. and Williams, J.B.W. (2001) The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, **16**, 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[39] Spitzer, R.L., Kroenke, K., Williams, J.B.W. and Löwe, B. (2006) A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, **166**, 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

[40] Velligan, D.I., Weiden, P.J., Sajatovic, M., Scott, J., Carpenter, D., Ross, R., *et al.* (2010) Strategies for Addressing Adherence Problems in Patients with Serious and Persistent Mental Illness: Recommendations from the Expert Consensus Guidelines. *Journal of Psychiatric Practice*, **16**, 306-324. https://doi.org/10.1097/01.pra.0000388626.98662.a0

[41] Green, C.A., Yarborough, B.J.H., Leo, M.C., Yarborough, M.T., Stumbo, S.P., Janoff, S.L., *et al.* (2015) The STRIDE Weight Loss and Lifestyle Intervention for Individuals Taking Antipsychotic Medications: A Randomized Trial. *American Journal of Psychiatry*, **172**, 71-81. https://doi.org/10.1176/appi.ajp.2014.14020173

[42] Bickman, L., Andrade, A.R. and Lambert, E.W. (2002) Dose Response in Child and Adolescent Mental Health Services. *Mental Health Services Research*, **4**, 57-70. https://doi.org/10.1023/a:1015210332175

[43] Garland, A.F., Haine, R.A. and Lewczyk Boxmeyer, C. (2007) Correlates of Adolescents' Satisfaction with Mental Health Services. *Mental Health Services Research*, **9**, 263-270.

[44] Paszke, A., Gross, S., Massa, F., *et al.* (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Annual Conference on Neural Information Processing Systems* 2019, *NeurIPS* 2019, Vancouver, 8-14 December 2019, 8024-8035.

[45] Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830.

[46] Bottou, L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent. In: *Proceedings of COMPSTAT'*2010, Physica-Verlag HD, 177-186. https://doi.org/10.1007/978-3-7908-2604-3_16

[47] Nguyen, L.M., Scheinberg, K., and Takáč, M. (2021) Inexact SARAH Algorithm for Stochastic Optimization. *Optimization Methods and Software*, **36**, 237-258. https://doi.org/10.1080/10556788.2020.1818081

[48] Srivastava, N., Hinton, G., Krizhevsky, A., *et al.* (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, **15**, 1929-1958.

[49] Ioffe, S. and Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning*, Lille, 6-11 July 2015, 448-456.

[50] Chen, Y.J., Yue, N., Slawski, M., and Rangwala, H. (2020) Asynchronous Online Federated Learning for Edge Devices with Non-IID Data. 2020 *IEEE International Conference on Big Data* (*Big Data*), Atlanta, 10-13 December 2020, 15-24. https://doi.org/10.1109/BigData50022.2020.9378161

[51] Lin, T., Kong, L., Stich, S.U. and Jaggi, M. (2020) Ensemble Distillation for Robust Model Fusion in Federated Learning. *Advances in Neural Information Processing Systems*, 6-12 December 2020, 2351-2363.

[52] Hsu, T.M., Qi, H. and Brown, M. (2019) Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. https://arxiv.org/abs/1909.06335

[53] Geyer, R.C., Klein, T. and Nabi, M. (2017) Differentially Private Federated Learning:

A Client Level Perspective. https://arxiv.org/abs/1712.07557

[54] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M. (2006) Our Data, Ourselves: Privacy via Distributed Noise Generation. In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Springer, 486-503. https://doi.org/10.1007/11761679_29

[55] Mironov, I. (2017) Rényi Differential Privacy. 2017 *IEEE* 30*th Computer Security Foundations Symposium* (*CSF*), Santa Barbara, 21-25 August 2017, 263-275. https://doi.org/10.1109/csf.2017.11

[56] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[57] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[58] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[59] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., *et al.* (2010) Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*, **21**, 128-138. https://doi.org/10.1097/ede.0b013e3181c30fb2

[60] Efron, B. and Tibshirani, R.J. (1994) An Introduction to the Bootstrap. CRC Press.

[61] Carpenter, J. and Bithell, J. (2000) Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians. *Statistics in Medicine*, **19**, 1141-1164. https://doi.org/10.1002/(sici)1097-0258(20000515)19:9<1141::aid-sim479>3.0.co;2-f

[62] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[63] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[64] Konečný, J., McMahan, H.B., Yu, F.X., *et al.* (2016) Federated Learning: Strategies for Improving Communication Efficiency. https://arxiv.org/abs/1610.05492

[65] Sattler, F., Wiedemann, S., Muller, K. and Samek, W. (2020) Robust and Communication-Efficient Federated Learning from Non-IID Data. *IEEE Transactions on Neural Networks and Learning Systems*, **31**, 3400-3413. https://doi.org/10.1109/tnnls.2019.2944481

[66] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**, 837-845. https://doi.org/10.2307/2531595

[67] Sun, X. and Xu, W. (2014) Fast Implementation of Delong's Algorithm for Comparing the Areas under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters*, **21**, 1389-1393. https://doi.org/10.1109/lsp.2014.2337313

[68] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., *et al.* (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, **17**, 261-272. https://doi.org/10.1038/s41592-019-0686-2

[69] Gaboardi, M., Haeberlen, A., Hsu, J., *et al.* (2016) Psi (Ψ): A Private Data Sharing Interface. https://arxiv.org/abs/1609.04340

[70] Vigod, S.N., Taylor, V.H., Fung, K., *et al.* (2015) Within-Year Readmission Risk for Psychiatric Disorders: Population-Based Cohort Study. *Psychiatric Services*, **66**, 1176-

1181.

[71] Lin, W.C., Zhang, J., Lake, A.J., *et al.* (2019) Early Readmission after Hospital Discharge among Patients with Mental Disorders in the United States. *Psychiatric Services*, **70**, 685-688.

[72] Collins, G.S. and Altman, D.G. (2012) Predicting the 10 Year Risk of Cardiovascular Disease in the United Kingdom: Independent and External Validation of an Updated Version of QRISK2. *BMJ*, **344**, e4181. https://doi.org/10.1136/bmj.e4181

[73] Steyerberg, E.W., Moons, K.G.M., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., *et al.* (2013) Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLOS Medicine*, **10**, e1001381. https://doi.org/10.1371/journal.pmed.1001381

[74] Nicholson, J., Krishnamurthy, R. and Lucas, P. (2021) Sharing Data for Public Health Research: A Systematic Review of Contextual Determinants. *International Journal of Environmental Research and Public Health*, **18**, 3217.

[75] Carter, P., Laurie, G.T. and Dixon-Woods, M. (2015) The Social Licence for Research: Why *care.data* Ran into Trouble. *Journal of Medical Ethics*, **41**, 404-409. https://doi.org/10.1136/medethics-2014-102374

[76] Shojania, K.G. and Forster, A.J. (2008) Hospital Mortality: When Failure Is Not a Good Measure of Success. *Canadian Medical Association Journal*, **179**, 153-157. https://doi.org/10.1503/cmaj.080010

[77] Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., *et al.* (2018) Differential Privacy: A Primer for a Non-Technical Audience. *Vanderbilt Journal of Entertainment & Technology Law*, **21**, 17.

[78] Gostin, L.O. and Hodge, J.G.H. (2002) Personal Privacy and Common Goods: A Framework for Balancing under the National Health Information Privacy Rule. *Minnesota Law Review*, **86**, Article No. 1439. https://doi.org/10.24926/265535.2691

[79] Mello, M.M., Francer, J.K., Wilenzick, M., Teden, P., Bierer, B.E. and Barnes, M. (2013) Preparing for Responsible Sharing of Clinical Trial Data. *New England Journal of Medicine*, **369**, 1651-1658. https://doi.org/10.1056/nejmhle1309073

[80] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S. (2016) Deep Learning for Visual Understanding: A Review. *Neurocomputing*, **187**, 27-48. https://doi.org/10.1016/j.neucom.2015.09.116

[81] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al.* (2017) Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. https://doi.org/10.1038/nature21056

[82] Zhao, Y., Li, M., Lai, L., *et al.* (2018) Federated Learning with Non-IID Data. https://arxiv.org/abs/1806.00582

[83] Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., *et al.* (2018) Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLOS Medicine*, **15**, e1002686. https://doi.org/10.1371/journal.pmed.1002686

[84] Oakden-Rayner, L., Dunnmon, J., Carneiro, G. and Re, C. (2020). Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, Toronto, 23-25 July 2020, 151-159. https://doi.org/10.1145/3368555.3384468

[85] Bonawitz, K., Eichner, H., Grieskamp, W., *et al.* (2019) Towards Federated Learning at Scale: System Design. *Proceedings of Machine Learning and Systems*, Vol. 1, 374-388.

[86] Lai, F., Dai, Y., Zhu, X., *et al.* (2021) FedScale: Benchmarking Model and System Per-

formance of Federated Learning. *Proceedings of the* 1*st Workshop on Distributed Machine Learning*, 25 October 2021, 1-3.

[87] Olfson, M., Wall, M., Wang, S., Crystal, S., Liu, S., Gerhard, T., *et al.* (2016) Short-term Suicide Risk after Psychiatric Hospital Discharge. *JAMA Psychiatry*, **73**, 1119-1126. https://doi.org/10.1001/jamapsychiatry.2016.2035

[88] Chung, D.T., Ryan, C.J., Hadzi-Pavlovic, D., Singh, S.P., Stanton, C. and Large, M.M. (2017) Suicide Rates after Discharge from Psychiatric Facilities: A Systematic Review and Meta-Analysis. *JAMA Psychiatry*, **74**, 694-702. https://doi.org/10.1001/jamapsychiatry.2017.1044

[89] So, J., Guler, B. and Avestimehr, A.S. (2021) Turbo-aggregate: Breaking the Quadratic Aggregation Barrier in Secure Federated Learning. *IEEE Journal on Selected Areas in Information Theory*, **2**, 479-489. https://doi.org/10.1109/jsait.2021.3054610

[90] Bell, J.H., Bonawitz, K.A., Gascón, A., Lepoint, T. and Raykova, M. (2020) Secure Single-Server Aggregation with (Poly)logarithmic Overhead. *Proceedings of the* 2020 *ACM SIGSAC Conference on Computer and Communications Security*, 9-13 November 2020, 1253-1269. https://doi.org/10.1145/3372297.3417885

[91] Zhuang, W., Gan, X., Wen, Y., Zhang, S. and Yi, S. (2021) Collaborative Unsupervised Visual Representation Learning from Decentralized Data. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 4912-4921. https://doi.org/10.1109/iccv48922.2021.00487

[92] Yao, D., Pan, W., Dai, Y., *et al.* (2019) FedLearn: Federated Machine Learning with Model Exchange. *Proceedings of the* 2019 *SIAM International Conference on Data Mining*, Calgary, 2-4 May 2019, 608-610.